医学情報処理演習第５回「データの分布と検定の概念」

中澤　港 (nminato@med.gunma-u.ac.jp)

2004 年 11 月 8 日

データの分布

ベルヌーイ試行と２項分布

１回の実験で事象 S か事象 F のどちらかが起こり，しかもそれらが起こる可能性が，$Pr(S) = p, Pr(F) = 1 - p = q$ で何回実験しても変わらないとき，これを「ベルヌーイ試行」という。ベルヌーイ試行では，事象 F は事象 S の余事象になっている。

例えば，不透明な袋に黒い玉と白い玉が 500 個ずつ入っていて，そこから中を見ないで１つの玉を取り出して色を記録して（事象 S は「玉の色が黒」，事象 F は「玉の色が白」）袋に戻す実験はベルヌーイ試行である（注：袋に戻さないと１回実験するごとに事象の生起確率が変わっていくのでベルヌーイ試行にならない。なお，サンプリングとみれば，これは復元抽出である）。

ベルヌーイ試行を n 回行って，S がちょうど k 回起こる確率は，$Pr(X = k) =_n C_k p^k q^{n-k}$ である。${}_nC_k$ は言うまでもなく n 個のものから k 個を取り出す組み合わせの数である。２項係数と呼ばれる。このような確率変数 X は，「２項分布に従う」といい，$X \sim B(n, p)$ と表す。$E(X) = np$, $V(X) = npq$ である。

２項分布のシミュレーション

正二十面体（各面には１から 20 までの数字が割り振られている）サイコロを n 回 $(n = 4, 10, 20, 50)$ 投げたときの，１から４までの目が出る回数を１試行と考えれば，これはベルヌーイ試行である。１回投げたときに１から４までの目が出る確率は 0.2 であるとして（＝母比率を 0.2 とする），試行 1000 セットの度数分布を描く R のプログラムは下記の通り。

```
times <- function(n) {
hit <- 0
dice <- as.integer(runif(n,1,21))
for (j in 1:n) { if (dice[j]<5) { hit <- hit+1 } }
return(hit)}

a <- c(4,10,20,50)
par(mfrow=c(2,2))
for (i in 1:4) {
nx <- a[i]
y <- c(1:1000)
for (k in 1:1000) { y[k] <- times(nx) }
barplot(table(y),main=paste("n=",nx))
}
```

２項分布の理論分布

この例で，各 n についての理論的な確率分布は，$Pr(X = k) =_n C_k 0.2^k 0.8^{n-k}$ なので，図を描くための R のプログラムは下記の通り。

```
a <- c(4,10,20,50)
par(mfrow=c(2,2))
for (i in 1:4) {
n <- a[i]
k <- 0
chk <- c(1:n+1)
while (k <= n) { chk[k+1] <- choose(n,k)*(0.2^k)*(0.8^(n-k)); k <- k+1 }
barplot(chk,main=paste("n=",n))
}
```

ただし，R には様々な確率分布についての関数があり，`choose(n,k)*(0.2^k)*(0.8^(n-k))` は `dbinom(k,n,0.2)` と同値である。このように，確率変数が取りうる各値に対して，その値をとる確率を与える関数を確率密度関数という。値が小さいほうからそれを全部足した値を与える関数（つまり，その確率変数の標本空間の下限から各値までの確率密度関数の定積分）を分布関数（あるいは確率母関数，累積確率密度関数）と呼ぶ。

正規分布

n が非常に大きい場合は，2 項分布 $B(n,p)$ の確率 $Pr(X = np + d)$ という値が，

$$\frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{d^2}{2npq}\right)$$

で近似できる。一般にこの極限（n を無限大に限りなく近づけた場合）である，

$$Pr(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

という形をもつ確率分布を正規分布と呼び，$N(\mu, \sigma^2)$ と書く。

$z = (x - \mu)/\sigma$ と置けば，

$$Pr(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

となる。これを標準正規分布と呼び，$N(0, 1)$ と書く。

R で標準正規分布の確率密度関数をプロットする方法は前回示したように

```
X <- 1:1000/100-5
plot(X,dnorm(X),type="l")
```

とすればよい（注：type=”l”は line の略で，線で繋ぐことを意味する）。標準正規分布の 97.5%点（その点より小さい値をとる確率の積分値が 0.975 になるような点）を得るには，`qnorm(0.975)` とすればよいし，-1.96 より小さな値をとる確率を得るには，`pnorm(-1.96)` とすればよい。

χ^2 分布

$X_1, X_2, ..., X_v$ が互いに独立に標準正規分布 $N(0, 1)$ に従うとき，

$$V = \sum_{i=1}^{v} {X_i}^2$$

の分布を自由度 v の χ^2 分布という。この分布の確率密度関数は，

$$f(x|v) = \frac{1}{2\Gamma(v/2)} \left(\frac{v}{2}\right)^{v/2-1} \exp\left(-\frac{v}{2}\right)$$

である。

なお，言うまでもないが，Γ はガンマ関数で，正の実数 α に対して，

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$$

であり，正の整数 α に対しては $\Gamma(\alpha) = (\alpha - 1)!$ である。

$E(x) = v$ であり，$V(x) = 2v$ である。自由度 1 の χ^2 分布をプロットするには，正規分布を描くときと同様に X を定義した後で，`plot(X,dchisq(X,1),type="l")` とすればよい。他の自由度のものを重ね描きするには，`lines(X,dchisq(X,2))` などとすればよい。自由度 1 の χ^2 分布の 95%点を得るには，`qchisq(0.95,1)` とすればよいし，3.84 より小さな値をとる確率を得るには，`pchisq(3.84,1)` とすればよい。

t 分布

標準正規分布に従う確率変数 U と，自由度 v の χ^2 分布 $\chi^2(v)$ に従う確率変数 V があり，それらが独立のとき，

$$T = U/\sqrt{V/v}$$

が従う分布のことをステューデントの t 分布という。この確率密度関数は

$$f(t) = \frac{\Gamma((v+1)/2)}{\sqrt{v}\Gamma(1/2)\Gamma(v/2)}\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}$$

である。これは，ステューデントというペンネームで論文を書いていたギネス社の技師ゴセット (Gosset WS) が初めて導いた分布である。

自由度 20 の t 分布の確率密度関数をプロットするには，正規分布を描くときと同様に X を定義した後で，`plot(X,dt(X,20),type="l")` とすればよい。これが標準正規分布より裾が長い分布であることを見るためには，続けて，`lines(X,dnorm(X),col="red")` とすればよい。また，自由度 20 の t 分布の 97.5%点を得るには，`qt(0.975,20)` とすればよいし，2 より小さな値をとる確率を得るには，`pt(2,20)` とすればよい。

分布の正規性の検定

高度な統計解析をするときには，データが正規分布する母集団からのサンプルであるという仮定を置くことが多いのだけれども，それを実際に確認することは難しいので，一般には，分布の正規性の検定を行うことが多い。考案者の名前から Shapiro-Wilk の検定と呼ばれるものが代表的である。

Shapiro-Wilk の検定の原理をざっと説明すると，$Z_i = (X_i - \mu)/\sigma$ とおけば，Z_i が帰無仮説「X が正規分布にしたがう」の下で $N(0,1)$ からの標本の順序統計量となり，$c(i) = E[Z(i)], d_{ij} = Cov(Z(i), Z(j))$ が母数に無関係な定数となるので，「$X(1) < X(2) < ... < X(n)$ の $c(1), c(2), ...c(n)$ への回帰が線形である」を帰無仮説として，そのモデルの下で σ の最良線形不偏推定量 $\hat{(}\sigma) = \sum_{i=1}^{n} a_i X(i)$ と $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ を用いて，$W = (k\hat{\sigma}^2)/S^2$ を検定統計量として検定するものである。k は $\sum_{i=1}^{n}(ka_i)^2 = 1$ より求められる。

R で数値型変数 X の分布が正規分布にフィットしているかどうかを検定するには，

```
shapiro.test(X)
```

とすればよい。変数 X のデータ数（ベクトルの要素数）は，3 から 5000 の間でなければならない。2 以下では分布を考える意味がなく，また，検定統計量 W の分布がモンテカルロシミュレーションによって得られたものであるためである。

正規性の検定には，もう 1 つ，ギアリー (Geary) の検定と呼ばれるものもある。現在のところ，R にはデフォルトでは入っていないようだが，左右対称な分布について，裾の長さを平均値のまわりの 1 次の絶対モーメントを 2 次のモーメントの平方根で割ったもので測ることにすると，その一致推定量 G が，

$$G = \frac{\sum_{i=1}^{n}|X_i - \bar{X}|}{\sqrt{n\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

となる。この G を用いて帰無仮説 H0:「データ X が正規分布からの標本」を検定することができる。対立仮説の下での分布が正規分布よりも裾が長い対称分布（例えば t 分布のような）であれば $G < g_0$ のとき帰無仮説を棄却する。g のパーセント点については，u_α を標準正規分布の 100α%点として，n が大きければ近似的に

$$g(\alpha; n) \simeq \sqrt{\frac{2}{\pi}} + u_\alpha\sqrt{1 - \frac{3}{\pi}}\frac{1}{\sqrt{n}}$$

で得られることがわかっている。そこで，R のプログラムとしては下枠内のようにして，例えば `G` が `g(0.05)` の値よりも小さければ，正規分布よりも 5%水準で有意に裾が長いとして帰無仮説を棄却すればよい。下枠内の関数定義をした後で，試しに `geary.test(rnorm(1000))` とすれば，ほぼ全ての試行で `G` は `g(0.05)` よりも大きくなるし，`geary.test(rt(1000,20))` とすれば，ほぼ全ての試行で `G` は `g(0.01)` よりも小さくなるであろう。

```
geary.test <- function(X) {
m.X <- mean(X)
l.X <- length(X)
G <- sum(abs(X-m.X))/sqrt(l.X*sum((X-m.X)^2))
g5 <- sqrt(2/pi)+qnorm(0.05)*sqrt(1-3/pi)/sqrt(l.X)
g1 <- sqrt(2/pi)+qnorm(0.01)*sqrt(1-3/pi)/sqrt(l.X)
cat("G=",G,"  /  g(0.05)=",g5,"  /  g(0.01)=",g1,"\n") }
```

作図の説明で触れた `qqnorm(X)` をしてみるのも，分布の正規性をチェックするにはいい方法である。正規分布に従っていれば直線に乗るはずであり，外れているときにどのように外れているかが見える。

検定と第一種，第二種の過誤

検定とは，帰無仮説の元で得られた統計量を，既知の確率分布をもつ量と見た場合に，その値よりも外れた値が得られる確率がどれほど小さいかを調べ，有意水準（5%とか 1%とか，分析者が決める）より小さければ，統計的に意味があることと捉え（統計的に有意である，という），帰無仮説がおかしいと判断して棄却するという意思決定を行うものである。

この意思決定が間違っていて，本当は帰無仮説が正しいのに，間違って帰無仮説を棄却してしまう確率は，有意水準と等しいので，その意味で，有意水準を第一種の過誤と（αエラーとも）呼ぶ（逆に，本当は帰無仮説が正しくないのに，その差を検出できず，有意でないと判断してしまう確率を，第二種の過誤と（βエラーとも）呼び，1 から第二種の過誤を引いた値が検出力になる）。

両側検定と片側検定

2 つの量的変数 X と Y の平均値の差の検定をする場合（平均値の差の検定についてはまた次回詳しく触れる），それぞれの母平均を μ_X，μ_Y と書けば，その推定量は $\mu_X = \mathrm{mean}(X) = \sum X/n$ と $\mu_Y = \mathrm{mean}(Y) = \sum Y/n$ となる。

両側検定では，帰無仮説 H0 : $\mu_X = \mu_Y$ に対して対立仮説（帰無仮説が棄却された場合に採択される仮説）H1 : $\mu_X \neq \mu_Y$ である。H1 を書き直すと，「$\mu_X > \mu_Y$ または $\mu_X < \mu_Y$」ということである。つまり，t_0 を「平均値の差を標準誤差で割った値」として求めると，t_0 が負になる場合も正になる場合もあるので，有意水準 5%で検定して有意になる場合というのは，t_0 が負で t 分布の下側 2.5%点より小さい場合と，t_0 が正で t 分布の上側 2.5%点（つまり 97.5%点）より大きい場合の両方を含む。t 分布は原点について対称なので，結局両側検定の場合は，上述のように差の絶対値を分子にして，t_0 の t 分布の上側確率[*1]を 2 倍すれば有意確率が得られることになる。

片側検定は，先験的に X と Y の間に大小関係が仮定できる場合に行い，例えば，X の方が Y より小さくなっているかどうかを検定したい場合なら，帰無仮説 H0 : $\mu_X \geq \mu_Y$ に対して対立仮説 H1 : $\mu_X < \mu_Y$ となる。この場合は，t_0 が正になる場合だけ考えればよい。有意水準 5%で検定して有意になるのは，t_0 が t 分布の上側 5%点（つまり 95%点）より大きい場合である。R で片側検定をしたい場合は，`alternative` という指定を追加する。例えば，$X > Y$ が対立仮説なら，`t.test(X,Y,alternative="greater")` とする。指定しなければ両側検定である。`alternative` に指定できる文字列は，`greater` の他には `less` と `two.sided` がある（指定しない場合は `two.sided` を指定したのと同じ意味，つまり両側検定になる）。

課題

片側検定をした方がいい例を想定せよ（思いつかなければ，今回も含めて，これまでの演習への意見・質問・感想でもよい）。

結果は配布する紙に学籍番号，氏名と共に自筆して提出すること。結果の提出をもって出席確認とする。

[*1] t 分布の確率密度関数を t_0 から無限大まで積分した値，即ち，t 分布の分布関数の t_0 のところの値を 1 から引いた値。R では `1-pt(t0, 自由度)`。